Homework 1
Runzhong Zhang rz2431

1. Overview

This assignment mainly contains four parts: (1) compute number of bytes served to every IP; (2) return top *K* IPs that were served the greatest number of bytes; (3) compute the number of bytes served for every IP in one hour; (4) compute the same statistics for a subnet.

Here we use the dataset "epa-http.txt".

2. Implement details

2.1 Question (1)

The first question is very simple. The data format looks like this:

141.243.1.172 [29:23:53:25] "GET /Software.html HTTP/1.0" 200 1497

This is the first row of dataset. By observing the data I find out that the first part is IP, the second part is time, then followed by request, return code and bytes. Each part is separated by "space". Therefore, if we use "split" function and get the first and final parts, then use "reduceByKey" to sum up the bytes for the same IP, then the code is almost finished.

However, there are some missing data in the dataset:

tanuki.twics.com [29:23:54:19] "GET /docs/OSWRCRA/general/hotline HTTP/1.0" 302 –

As we can see, the number of bytes is missing, so there will be a bug if we simply convert the last part into "int" format. Therefore, we should pre-process the dataset by using "filter" function.

2.2 Question (2)

Based on the question (1), the second part is also very easy. I sort the result from question (1) with "sortBy" function.

2.3 Question (3)

Here we should divide the dataset based on hour, which is contained in second part of data. The second part looks like this: [day : hour : minute : second]. Therefore, if we split the data by ":", then we can get the hour.

2.4 Question (4)

According to the announcement on coursework, for non-digit IP addresses we should aggregate them by the first two words, and for digit IP addresses we should aggregate them by the first three digits. Therefore, my code first divide the dataset into two parts (digit IP and non-digit IP), and process them separately using "map" and "reduceByKey" function.

3. Results

3.1 Question (1)

For the readers' convenience, I use "take" function to show only the first 15 results:

[('141.243.1.172', 1634402), ('query2.lycos.cs.cmu.edu', '1325'), ('140.112.68.165', 7811), ('dd15-032.compuserve.com', 12898), ('freenet2.carleton.ca', '15173'), ('ix-mia5-17.ix.netcom.com',

42461), ('hmu4.cs.auckland.ac.nz', 257009), ('131.215.67.47', 32988), ('www-c1.proxy.aol.com', 205533), ('bettong.client.uq.oz.au', 177058), ('flaxman-q950.uoregon.edu', 32988), ('161.122.12.78', 3008684), ('137.132.52.66', 2235), ('playful.mnsinc.com', 2121), ('archives.math.utk.edu', 0)]

3.2 Question (2)

For the readers' convenience, I use "take" function to show only the first 15 results:

[('piankhi.cs.hamptonu.edu', 7267751), ('e659229.boeing.com', 5260561), ('139.121.98.45', 5041738), ('ws13dgadrv.er.usgs.gov', 4716720), ('slcmodem1-p1-14.intele.net', 4453807), ('www-c5.proxy.aol.com', 4435337), ('so.scsnet.com', 4420061), ('keyhole.es.dupont.com', 4005003), ('203.251.228.110', 3785626), ('cnts4p16.uwaterloo.ca', 3636398), ('rac3.wam.umd.edu', 3590760), ('dialin30.annex1.radix.net', 3405626), ('155.84.92.3', 3353172), ('198.102.67.27', 3182052), ('piweba5y.prodigy.com', 3084168)]

3.3 Question (3)

For the readers' convenience, I only show the data in "05 hour":

05 hour (the second day):

[('slip137-11.pt.uk.ibm.net', 54478), ('ngriffin.itc.gu.edu.au', 126062), ('ix-stp-fl1-20.ix.netcom.com', 699567), ('pc10.kcl.fi', 11747), ('161.122.12.78', '557056'), ('epsongw3.epson.co.jp', 2305845), ('volve20.vol.it', 10110), ('comnet2.ksc.net.th', 57296), ('pc032058.eeng.liv.ac.uk', 123374), ('volve15.vol.it', 608216), ('hcollinson.elsevier.co.uk', 64731), ('ppp017.st.rim.or.jp', 17578), ('193.145.151.115', 70986), ('141.243.1.187', 11747), ('168.154.26.10', 11747), ('slip-2.tizeta.it', 14412), ('vifa1.freenet.victoria.bc.ca', 6864), ('203.249.9.64', '1045'), ('141.243.1.186', 11747), ('204.62.245.32', 24842), ('202.46.240.10', 20658), ('oahu-81.u.aloha.net', '2624'), ('192.146.118.133', 84319), ('130.158.72.126', 40575), ('oslo102.telepost.no', 36823), ('ppp006.st.rim.or.jp', 16710), ('cybele.rug.ac.be', '4889'), ('152.165.111.120', 7758), ('proxy.iijnet.or.jp', 11747), ('192.228.164.77', 7758), ('ppp226.gol.com', 181424), ('pc27.iuv.uni-bremen.de', 10137), ('library.wustl.edu', '4889')]

3.4 Question (4)

For the readers' convenience, I only show the first 5 results for two types of IP:

[('140.112.68', 7811), ('202.32.50', 159922), ('149.159.22', 67951), ('198.69.241', 34462), ('202.96.29', 10296)]

[('query2.lycos', '1325'), ('tanuki.twics', 60936), ('dd15-032.compuserve', 12898), ('freenet2.carleton', '15173'), ('ix-knx-tn1-22.ix', 14450)]

4. Summary

My codes successfully solve all the problems. If you meet any issues when running the code, please contract me at rz2431@columbia.edu. Thanks for reading my assignment.